Assignment Subjective questions

Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

The optimal value of alpha for Ridge and Lasso Regression are: 10 and 0.001.

As the value of alpha increases for lasso, the model complexity reduces and the model tries to make more coefficients as zeros. So here, when we change the value to 0.002, since the value of alpha is very small, we don't see much impact, but yes we can see the coefficients being pulled much closer to zero.

As the value of alpha increases for ridge, the model complexity reduces in the case of ridge. But high values of alpha leads to underfitting. RSS value increases and the coefficients tends to zero, though not exact zero.

Lasso top features after doubling the value of alpha.

	Variable	Coeff	AbsCoeff
24	MSZoning_RL	3.156500e-01	0.3156
8	GrLivArea	2.691916e-01	0.2692
25	MSZoning_RM	2.223480e-01	0.2223
17	house_age	-2.071674e-01	0.2072
22	MSZoning_FV	1.571870e-01	0.1572
42	OverallQual_8	1.381013e-01	0.1381
0	LotArea	1.185736e-01	0.1186
43	OverallQual_9	1.140619e-01	0.1141
23	MSZoning_RH	8.996461e-02	0.0900
50	OverallCond_7	8.182816e-02	0.0818
73	GarageType_Attchd	8.021114e-02	0.0802

Ridge top features after doubling the value of alpha.

	Variable	Coeff	AbsCoeff
38	MSZoning_RL	0.201177	0.2012
26	house_age	-0.147684	0.1477
10	GrLivArea	0.127488	0.1275
39	MSZoning_RM	0.126035	0.1260
69	OverallQual_8	0.116102	0.1161
1	LotArea	0.112708	0.1127
36	MSZoning_FV	0.104390	0.1044
8	2ndFlrSF	0.101540	0.1015
70	OverallQual_9	0.093194	0.0932
106	BsmtQual_TA	-0.081249	0.0812
13	FullBath	0.081078	0.0811

Question 2

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer:

We determine the optimal value of lambda for ridge and lasso by comparing each cross-validation score. Here in my model, the ridge regression produces better score than lasso (though a very small difference) hence proving that shrinking the coefficients gives better results compared to making the coefficients zero as is the case with Lasso. But since the scores in my case is not much different, and Lasso is eliminating less important features, I would prefer Lasso over Ridge. Also, higher value of lambda in ridge leads to underfitting. Hence Lasso is a better option to pick for predictions.

Question 3

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:

The five most Important predictor variables after removing the initial five important variables are:

	Variable	Coeff	AbsCoeff
6	2ndFlrSF	1.921566e-01	0.1922
5	1stFlrSF	1.855009e-01	0.1855
37	OverallQual_8	1.782115e-01	0.1782
68	GarageType_Attchd	1.620972e-01	0.1621
38	OverallQual_9	1.408393e-01	0.1408

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

We can say the model is robust and generalisable when it is simple. Not too less features and not too many features should be considered. Less features leads to high bias and low variance. This leads to underfitting of the model. Large number of features can lead to overfitting (low bias and high variance). Thus bias-variance trade-off should be there. These Generic models are bound to perform better on unseen datasets.

There should be a trade-off between Accuracy and Generalisation also.

The model accuracy on training dataset is quite high for complex model compared to simple model. However, if we extrapolate the fitted lines for both the models, the complex model fails miserably while predicting on the test data.